3) EP 00 11 5227.1

(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) N° de publication : **2 768 547**
(à n'utiliser que pour les commandes de reproduction)

(21) N° d'enregistrement national : **97 11643**

(51) Int Cl^6 : G 10 L 3/02, G 10 L 7/06

(12) DEMANDE DE BREVET D'INVENTION A1

(22) Date de dépôt : 18.09.97.

(30) Priorité :

(43) Date de mise à la disposition du public de la demande : 19.03.99 Bulletin 99/11.

(56) Liste des documents cités dans le rapport de recherche préliminaire : *Se reporter à la fin du présent fascicule*

(60) Références à d'autres documents nationaux apparentés :

(71) Demandeur(s) : *MATRA COMMUNICATION SOCIETE PAR ACTIONS SIMPLIFIEE* — FR.

(72) Inventeur(s) : LOCKWOOD PHILIP et LUBIARZ STEPHANE.

(73) Titulaire(s) :

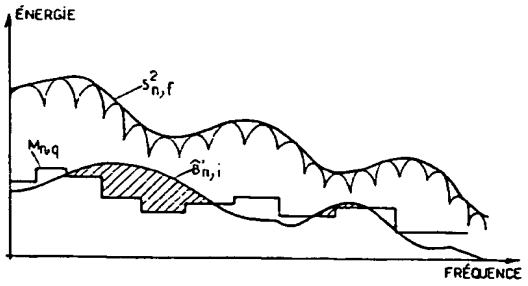
(74) Mandataire(s) : CABINET PLASSERAUD.

(54) PROCEDE DE DEBRUITAGE D'UN SIGNAL DE PAROLE NUMERIQUE.

(57) On effectue une soustraction spectrale comportant: une première étape de soustraction dans laquelle on tient compte d'estimations majorées

$(\hat{B}'_{n,i})$

des composantes spectrale du bruit, de manière à obtenir des composantes spectrales ($S^2_{n,i}$) d'un premier signal débruité; le calcul d'une courbe de masquage ($M_{n,q}$) en appliquant un modèle de perception auditive à partir des composantes spectrales du premier signal débruité; et une seconde étape de soustraction dans laquelle on soustrait respectivement, de chaque composante spectrale du signal de parole sur la trame, une quantité dépendant de paramètres incluant un écart entre l'estimation majorée de la composante spectrale correspondante du bruit et la courbe de masquage calculée. On applique au résultat de la soustraction spectrale une transformation vers le domaine temporel pour construire un signal de parole débruité.



FR 2 768 547 - A1

PROCEDE DE DEBRUITAGE D'UN SIGNAL DE PAROLE NUMERIQUE

La présente invention concerne les techniques numériques de débruitage de signaux de parole. Elle concerne plus particulièrement le débruitage par soustraction spectrale non linéaire.

Du fait de la généralisation des nouvelles formes de communication, en particulier des téléphones mobiles, les communications se font de plus en plus dans des ambiances fortement bruitées. Le bruit, additionné à la parole, a alors tendance à perturber les communications en empêchant une compression optimale du signal de parole et en créant un bruit de fond non naturel. D'autre part, le bruit rend difficile et fatigante la compréhension du message parlé.

De nombreux algorithmes ont été étudiés pour essayer de diminuer les effets du bruit dans une communication. S. F. Boll («Suppression of acoustic noise in speech using spectral subtraction », IEEE Trans. on Acoustics, Speech and Signal Processing », Vol. ASSP-27, n° 2, avril 1979) a proposé un algorithme basé sur la soustraction spectrale. Cette technique consiste à estimer le spectre du bruit pendant les phases de silence et à le soustraire du signal reçu. Elle permet une réduction du niveau de bruit reçu. Son principal défaut est de créer un bruit musical particulièrement gênant, car non naturel.

Ces travaux, repris et améliorés par D. B. Paul (« The spectral enveloppe estimation vocoder », IEEE Trans. on Acoustics, Speech and Signal Processing », Vol. ASSP-29, n° 4, août 1981) et par P. Lockwood et J. Boudy (« Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars », Speech Communication, Vol. 11, juin 1992, pages 215-228, et EP-A-0 534 837) ont permis de

diminuer sensiblement le niveau de bruit tout en lui conservant un caractère naturel. De plus, cette contribution a eu le mérite d'incorporer pour la première fois le principe de masquage dans le calcul du filtre de débruitage. A partir de cette idée, une première tentative a été faite par S. Nandkumar et J. H. L. Hansen (« Speech enhancement on a new set of auditory constrained parameters », Proc. ICASSP 94, pages I.1-I.4) pour utiliser dans la soustraction spectrale des courbes de masquage calculées explicitement. Malgré les résultats décevants de cette technique, cette contribution a eu le mérite de mettre l'accent sur l'importance de ne pas dénaturer le signal de parole pendant le débruitage.

D'autres méthodes basées sur la décomposition du signal de parole en valeurs singulières, et donc sur une projection du signal de parole dans un espace plus réduit, ont été étudiées par Bart De Moore (« The singular value decomposition and long and short spaces of noisy matrices », IEEE Trans. on Signal Processing, Vol. 41, n° 9, septembre 1993, pages 2826-2838) et par S. H. Jensen et al (« Reduction of broad-band noise in speech by truncated QSVD », IEEE Trans. on Speech and Audio Processing, Vol. 3, n° 6, novembre 1995). Le principe de cette technique est de considérer le signal de parole et le signal de bruit comme totalement décorrélés, et de considérer que le signal de parole a une prédictibilité suffisante pour être prédit à partir d'un jeu restreint de paramètres. Cette technique permet d'obtenir un débruitage acceptable pour des signaux fortement voisés, mais dénature totalement le signal de parole. Face à un bruit relativement cohérent, tel que celui provoqué par le contact de pneus de voitures ou le cliquetis d'un moteur, le bruit peut s'avérer plus facilement prédictible que le signal de parole non voisé. On a alors tendance à projeter le signal de parole dans

une partie de l'espace vectoriel du bruit. La méthode ne tient pas compte du signal de parole, en particulier des zones de parole non voisée où la prédictibilité est réduite. De plus, prédire le signal de parole à partir d'un jeu de paramètres réduit ne permet pas de prendre en compte toute la richesse intrinsèque de la parole. On comprend ici les limites de techniques basées uniquement sur des considérations mathématiques en oubliant le caractère particulier de la parole.

D'autres techniques enfin sont basées sur des critères de cohérence. La fonction de cohérence est particulièrement bien développée par J. A. Cadzow et O. M. Solomon (« Linear modeling and the coherence function », IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-35, n° 1, janvier 1987, pages 19-28), et son application au débruitage a été étudiée par R. Le Bouquin (« Enhancement of noisy speech signals : application to mobile radio communications », Speech Communication, Vol. 18, pages 3-19). Cette méthode se base sur le fait que le signal de parole a une cohérence nettement plus importante que le bruit à condition d'utiliser plusieurs canaux indépendants. Les résultats obtenus semblent être assez encourageants. Mais malheureusement, cette technique impose d'avoir plusieurs sources de prise de son, ce qui n'est pas toujours réalisé.

Un but principal de la présente invention est de proposer une nouvelle technique de débruitage qui prenne en compte les caractéristiques de perception de la parole par l'oreille humaine, permettant ainsi un débruitage efficace sans détériorer la perception de la parole.

L'invention propose ainsi un procédé de débruitage d'un signal de parole numérique traité par trames successives, dans lequel :

- on calcule des composantes spectrales du signal

de parole sur chaque trame ;

- on calcule pour chaque trame des estimations majorées de composantes spectrales du bruit compris dans le signal de parole ;

- on effectue une soustraction spectrale comportant au moins une première étape de soustraction dans laquelle on soustrait respectivement, de chaque composante spectrale du signal de parole sur la trame, une première quantité dépendant de paramètres incluant l'estimation majorée de la composante spectrale correspondante du bruit pour ladite trame, de manière à obtenir des composantes spectrales d'un premier signal débruité ; et

- on applique au résultat de la soustraction spectrale une transformation vers le domaine temporel pour construire un signal de parole débruité.

Selon l'invention, la soustraction spectrale comporte en outre les étapes suivantes :

- le calcul d'une courbe de masquage en appliquant un modèle de perception auditive à partir des composantes spectrales du premier signal débruité ;

- la comparaison des estimations majorées des composantes spectrales du bruit pour la trame à la courbe de masquage calculée ; et

- une seconde étape de soustraction dans laquelle on soustrait respectivement, de chaque composante spectrale du signal de parole sur la trame, une seconde quantité dépendant de paramètres incluant un écart entre l'estimation majorée de la composante spectrale correspondante du bruit et la courbe de masquage calculée.

La seconde quantité soustraite peut notamment être limitée à la fraction de l'estimation majorée de la composante spectrale correspondante du bruit qui dépasse la courbe de masquage. Cette façon de procéder repose sur l'observation qu'il suffit de débruiter les fréquences de bruit audibles. A contrario, il ne sert à rien d'éliminer

du bruit qui est masqué par de la parole.

La surestimation de l'enveloppe spectrale du bruit est généralement souhaitable pour que l'estimation majorée ainsi obtenue soit robuste aux brusques variations du bruit. Néanmoins, cette surestimation a habituellement l'inconvénient de distordre le signal de parole lorsqu'elle devient trop importante. Ceci a pour effet d'affecter le caractère voisé du signal de parole en supprimant une partie de sa prédictibilité. Cet inconvénient est très gênant dans les conditions de la téléphonie, car c'est pendant les zones de voisement que le signal de parole est alors le plus énergétique. En limitant la quantité soustraite lorsque la totalité ou une partie d'une composante fréquentielle du bruit surestimé s'avère être masquée par la parole, l'invention permet d'atténuer fortement cet inconvénient.

D'autres particularités et avantages de la présente invention apparaîtront dans la description ci-après d'exemples de réalisation non limitatifs, en référence aux dessins annexés, dans lesquels :

- la figure 1 est un schéma synoptique d'un système de débruitage mettant en œuvre la présente invention ;

- les figures 2 et 3 sont des organigrammes de procédures utilisées par un détecteur d'activité vocale du système de la figure 1 ;

- la figure 4 est un diagramme représentant les états d'un automate de détection d'activité vocale ;

- la figure 5 est un graphique illustrant les variations d'un degré d'activité vocale ;

- la figure 6 est un schéma synoptique d'un module de surestimation du bruit du système de la figure 1 ;

- la figure 7 est un graphique illustrant le calcul d'une courbe de masquage ;

- la figure 8 est un graphique illustrant l'exploitation des courbes de masquage dans le système de la figure 1 ;

- la figure 9 est un schéma synoptique d'un autre système de débruitage mettant en œuvre la présente invention ;

- la figure 10 est un graphique illustrant une méthode d'analyse harmonique utilisable dans un procédé selon l'invention ; et

- la figure 11 montre partiellement une variante du schéma synoptique de la figure 9.

Le système de débruitage représenté sur la figure 1 traite un signal numérique de parole s. Un module de fenêtrage 10 met ce signal s sous forme de fenêtres ou trames successives, constituées chacune d'un nombre N d'échantillons de signal numérique. De façon classique, ces trames peuvent présenter des recouvrements mutuels. Dans la suite de la présente description, on considérera, sans que ceci soit limitatif, que les trames sont constituées de N=256 échantillons à une fréquence d'échantillonnage F_e de 8 kHz, avec une pondération de Hamming dans chaque fenêtre, et des recouvrements de 50% entre fenêtres consécutives.

La trame de signal est transformée dans le domaine fréquentiel par un module 11 appliquant un algorithme classique de transformée de Fourier rapide (TFR) pour calculer le module du spectre du signal. Le module 11 délivre alors un ensemble de N=256 composantes fréquentielles du signal de parole, notées $S_{n,f}$, où n désigne le numéro de la trame courante, et f une fréquence du spectre discret. Du fait des propriétés des signaux numériques dans le domaine fréquentiel, seuls les N/2=128 premiers échantillons sont utilisés.

Pour calculer les estimations du bruit contenu dans le signal s, on n'utilise pas la résolution fréquentielle disponible en sortie de la transformée de Fourier rapide, mais une résolution plus faible, déterminée par un nombre I de bandes de fréquences couvrant la bande $[0,F_e/2]$ du signal. Chaque bande i ($1 \leq i \leq I$) s'étend entre une fréquence inférieure f(i-1) et une fréquence supérieure f(i), avec f(0)=0, et $f(I)=F_e/2$. Ce découpage en bandes de fréquences peut être uniforme ($f(i)-f(i-1)=F_e/2I$). Il peut également être non uniforme (par exemple selon une échelle de barks). Un module 12 calcule les moyennes respectives des composantes spectrales $S_{n,f}$ du signal de parole par bandes, par exemple par une pondération uniforme telle que :

$$S_{n,i} = \frac{1}{f(i) - f(i-1)} \sum_{f \in [f(i-1), f(i)[} S_{n,f} \qquad (1)$$

Ce moyennage diminue les fluctuations entre les bandes en moyennant les contributions du bruit dans ces bandes, ce qui diminuera la variance de l'estimateur de bruit. En outre, ce moyennage permet une forte diminution de la complexité du système.

Les composantes spectrales moyennées $S_{n,i}$ sont adressées à un module 15 de détection d'activité vocale et à un module 16 d'estimation du bruit. Ces deux modules 15, 16 fonctionnent conjointement, en ce sens que des degrés d'activité vocale $\gamma_{n,i}$ mesurés pour les différentes bandes par le module 15 sont utilisés par le module 16 pour estimer l'énergie à long terme du bruit dans les différentes bandes, tandis que ces estimations à long terme $\hat{B}_{n,i}$ sont utilisées par le module 15 pour procéder à un débruitage a priori du signal de parole dans les

différentes bandes pour déterminer les degrés d'activité vocale $\gamma_{n,i}$.

Le fonctionnement des modules 15 et 16 peut correspondre aux organigrammes représentés sur les figures 2 et 3.

Aux étapes 17 à 20, le module 15 procède au débruitage a priori du signal de parole dans les différentes bandes i pour la trame de signal n. Ce débruitage a priori est effectué selon un processus classique de soustraction spectrale non linéaire à partir d'estimations du bruit obtenues lors d'une ou plusieurs trames précédentes. A l'étape 17, le module 15 calcule, avec la résolution des bandes i, la réponse en fréquence $Hp_{n,i}$ du filtre de débruitage a priori, selon la formule :

$$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau1,i} \cdot \hat{B}_{n-\tau1,i}}{S_{n-\tau2,i}} \qquad (2)$$

où $\tau1$ et $\tau2$ sont des retards exprimés en nombre de trames ($\tau1 \geq 1$, $\tau2 \geq 0$), et $\alpha'_{n,i}$ est un coefficient de surestimation du bruit dont la détermination sera expliquée plus loin. Le retard $\tau1$ peut être fixe (par exemple $\tau1=1$) ou variable. Il est d'autant plus faible qu'on est confiant dans la détection d'activité vocale.

Aux étapes 18 à 20, les composantes spectrales $\hat{E}p_{n,i}$ sont calculées selon :

$$\hat{E}p_{n,i} = \max\left\{Hp_{n,i} \cdot S_{n,i} \, , \, \beta p_i \cdot \hat{B}_{n-\tau1,i}\right\} \qquad (3)$$

où βp_i est un coefficient de plancher proche de 0, servant classiquement à éviter que le spectre du signal débruité prenne des valeurs négatives ou trop faibles qui provoqueraient un bruit musical.

Les étapes 17 à 20 consistent donc essentiellement à soustraire du spectre du signal une estimation, majorée par le coefficient $\alpha'_{n-1,i}$, du spectre du bruit estimé a priori.

A l'étape 21, le module 15 calcule l'énergie du signal débruité a priori dans les différentes bandes i pour la trame n : $E_{n,i} = \hat{E}\rho^2_{n,i}$. Il calcule aussi une moyenne globale $E_{n,0}$ de l'énergie du signal débruité a priori, par une somme des énergies par bande $E_{n,i}$, pondérée par les largeurs de ces bandes. Dans les notations ci-dessous, l'indice i=0 sera utilisé pour désigner la bande globale du signal.

Aux étapes 22 et 23, le module 15 calcule, pour chaque bande i ($0 \leq i \leq I$), une grandeur $\Delta E_{n,i}$ représentant la variation à court terme de l'énergie du signal débruité dans la bande i, ainsi qu'une valeur à long terme $\bar{E}_{n,i}$ de l'énergie du signal débruité dans la bande i. La grandeur $\Delta E_{n,i}$ peut être calculée par une formule simplifiée de dérivation : $\Delta E_{n,i} = \left| \frac{E_{n-4,i} + E_{n-3,i} - E_{n-1,i} - E_{n,i}}{10} \right|$. Quant à l'énergie à long terme $\bar{E}_{n,i}$, elle peut être calculée à l'aide d'un facteur d'oubli B1 tel que $0<B1<1$, à savoir $\bar{E}_{n,i} = B1 \cdot \bar{E}_{n-1,i} + (1-B1) \cdot E_{n,i}$.

Après avoir calculé les énergies $E_{n,i}$ du signal débruité, ses variations à court terme $\Delta E_{n,i}$ et ses valeurs à long terme $\bar{E}_{n,i}$ de la manière indiquée sur la figure 2, le module 15 calcule, pour chaque bande i ($0 \leq i \leq I$), une valeur ρ_i représentative de l'évolution de

l'énergie du signal débruité. Ce calcul est effectué aux étapes 25 à 36 de la figure 3, exécutées pour chaque bande i entre i=0 et i=I. Ce calcul fait appel à un estimateur à long terme de l'enveloppe du bruit ba_i, à un estimateur interne bi_i et à un compteur de trames bruitées b_i.

A l'étape 25, la grandeur $\Delta E_{n,i}$ est comparée à un seuil $\varepsilon 1$. Si le seuil $\varepsilon 1$ n'est pas atteint, le compteur b_i est incrémenté d'une unité à l'étape 26. A l'étape 27, l'estimateur à long terme ba_i est comparé à la valeur de l'énergie lissée $\bar{E}_{n,i}$. Si $ba_i \geq \bar{E}_{n,i}$, l'estimateur ba_i est pris égal à la valeur lissée $\bar{E}_{n,i}$ à l'étape 28, et le compteur b_i est remis à zéro. La grandeur ρ_i, qui est prise égale au rapport $ba_i/\bar{E}_{n,i}$ (étape 36), est alors égale à 1.

Si l'étape 27 montre que $ba_i < \bar{E}_{n,i}$, le compteur b_i est comparé à une valeur limite bmax à l'étape 29. Si b_i>bmax, le signal est considéré comme trop stationnaire pour supporter de l'activité vocale. L'étape 28 précitée, qui revient à considérer que la trame ne comporte que du bruit, est alors exécutée. Si $b_i \leq$bmax à l'étape 29, l'estimateur interne bi_i est calculé à l'étape 33 selon :

$$bi_i = (1-Bm) \cdot \bar{E}_{n,i} + Bm \cdot ba_i \qquad (4)$$

Dans cette formule, Bm représente un coefficient de mise à jour compris entre 0,90 et 1. Sa valeur diffère selon l'état d'un automate de détection d'activité vocale (étapes 30 à 32). Cet état δ_{n-1} est celui déterminé lors du traitement de la trame précédente. Si l'automate est dans un état de détection de parole (δ_{n-1}=2 à l'étape 30),

le coefficient Bm prend une valeur Bmp très proche de 1 pour que l'estimateur du bruit soit très faiblement mis à jour en présence de parole. Dans le cas contraire, le coefficient Bm prend une valeur Bms plus faible, pour permettre une mise à jour plus significative de l'estimateur de bruit en phase de silence. A l'étape 34, l'écart ba_i-bi_i entre l'estimateur à long terme et l'estimateur interne du bruit est comparé à un seuil ε2. Si le seuil ε2 n'est pas atteint, l'estimateur à long terme ba_i est mis à jour avec la valeur de l'estimateur interne bi_i à l'étape 35. Sinon, l'estimateur à long terme ba_i reste inchangé. On évite ainsi que de brutales variations dues à un signal de parole conduisent à une mise à jour de l'estimateur de bruit.

Après avoir obtenu les grandeurs ρ_i, le module 15 procède aux décisions d'activité vocale à l'étape 37. Le module 15 met d'abord à jour l'état de l'automate de détection selon la grandeur ρ_0 calculée pour l'ensemble de la bande du signal. Le nouvel état δ_n de l'automate dépend de l'état précédent δ_{n-1} et de ρ_0, de la manière représentée sur la figure 4.

Quatre états sont possibles : $\delta=0$ détecte le silence, ou absence de parole ; $\delta=2$ détecte la présence d'une activité vocale ; et les états $\delta=1$ et $\delta=3$ sont des états intermédiaires de montée et de descente. Lorsque l'automate est dans l'état de silence ($\delta_{n-1}=0$), il y reste si ρ_0 ne dépasse pas un premier seuil SE1, et il passe dans l'état de montée dans le cas contraire. Dans l'état de montée ($\delta_{n-1}=1$), il revient dans l'état de silence si

ρ_0 est plus petit que le seuil SE1, il passe dans l'état de parole si ρ_0 est plus grand qu'un second seuil SE2 plus grand que le seuil SE1, et il reste dans l'état de montée si $SE1 \leq \rho_0 \leq SE2$. Lorsque l'automate est dans l'état de parole ($\delta_{n-1}=2$), il y reste si ρ_0 dépasse un troisième seuil SE3 plus petit que le seuil SE2, et il passe dans l'état de descente dans le cas contraire. Dans l'état de descente ($\delta_{n-1}=3$), l'automate revient dans l'état de parole si ρ_0 est plus grand que le seuil SE2, il revient dans l'état de silence si ρ_0 est en deçà d'un quatrième seuil SE4 plus petit que le seuil SE2, et il reste dans l'état de descente si $SE4 \leq \rho_0 \leq SE2$.

A l'étape 37, le module 15 calcule également les degrés d'activité vocale $\gamma_{n,i}$ dans chaque bande $i \geq 1$. Ce degré $\gamma_{n,i}$ est de préférence un paramètre non binaire, c'est-à-dire que la fonction $\gamma_{n,i}=g(\rho_i)$ est une fonction variant continûment entre 0 et 1 en fonction des valeurs prises par la grandeur ρ_i. Cette fonction a par exemple l'allure représentée sur la figure 5.

Le module 16 calcule les estimations du bruit par bande, qui seront utilisées dans le processus de débruitage, en utilisant les valeurs successives des composantes $S_{n,i}$ et des degrés d'activité vocale $\gamma_{n,i}$. Ceci correspond aux étapes 40 à 42 de la figure 3. A l'étape 40, on détermine si l'automate de détection d'activité vocale vient de passer de l'état de montée à l'état de parole. Dans l'affirmative, les deux dernières estimations $\hat{B}_{n-1,i}$ et $\hat{B}_{n-2,i}$ précédemment calculées pour

chaque bande $i \geq 1$ sont corrigées conformément à la valeur de l'estimation précédente $\hat{B}_{n-3,i}$. Cette correction est effectuée pour tenir compte du fait que, dans la phase de montée ($\delta=1$), les estimations à long terme de l'énergie du bruit dans le processus de détection d'activité vocale (étapes 30 à 33) ont pu être calculées comme si le signal ne comportait que du bruit (Bm=Bms), de sorte qu'elles risquent d'être entachées d'erreur.

A l'étape 42, le module 16 met à jour les estimations du bruit par bande selon les formules :

$$\tilde{B}_{n,i} = \lambda_B \cdot \hat{B}_{n-1,i} + (1-\lambda_B) \cdot S_{n,i} \qquad (5)$$

$$\hat{B}_{n,i} = \gamma_{n,i} \cdot \hat{B}_{n-1,i} + (1-\gamma_{n,i}) \cdot \tilde{B}_{n,i} \qquad (6)$$

où λ_B désigne un facteur d'oubli tel que $0<\lambda_B<1$. La formule (6) met en évidence la prise en compte du degré d'activité vocale non binaire $\gamma_{n,i}$.

Comme indiqué précédemment, les estimations à long terme du bruit $\hat{B}_{n,i}$ font l'objet d'une surestimation, par un module 45 (figure 1), avant de procéder au débruitage par soustraction spectrale non linéaire. Le module 45 calcule le coefficient de surestimation $\alpha'_{n,i}$ précédemment évoqué, ainsi qu'une estimation majorée $\hat{B}'_{n,i}$ qui correspond essentiellement à $\alpha'_{n,i} \cdot \hat{B}_{n,i}$.

L'organisation du module de surestimation 45 est représentée sur la figure 6. L'estimation majorée $\hat{B}'_{n,i}$ est obtenue en combinant l'estimation à long terme $\hat{B}_{n,i}$ et une mesure $\Delta B^{max}_{n,i}$ de la variabilité de la composante du bruit

dans la bande i autour de son estimation à long terme. Dans l'exemple considéré, cette combinaison est, pour l'essentiel, une simple somme réalisée par un additionneur 46. Ce pourrait également être une somme pondérée.

Le coefficient de surestimation $\alpha'_{n,i}$ est égal au rapport entre la somme $\hat{B}_{n,i} + \Delta B^{max}_{n,i}$ délivrée par l'additionneur 46 et l'estimation à long terme retardée $\hat{B}_{n-\tau 3,i}$ (diviseur 47), plafonné à une valeur limite α_{max}, par exemple $\alpha_{max}=4$ (bloc 48). Le retard $\tau 3$ sert à corriger le cas échéant, dans les phases de montée ($\delta=1$), la valeur du coefficient de surestimation $\alpha'_{n,i}$, avant que les estimations à long terme aient été corrigées par les étapes 40 et 41 de la figure 3 (par exemple $\tau 3=3$).

L'estimation majorée $\hat{B}'_{n,i}$ est finalement prise égale à $\alpha'_{n,i} \cdot \hat{B}_{n-\tau 3,i}$ (multiplieur 49).

La mesure $\Delta B^{max}_{n,i}$ de la variabilité du bruit reflète la variance de l'estimateur de bruit. Elle est obtenue en fonction des valeurs de $S_{n,i}$ et de $\hat{B}_{n,i}$ calculées pour un certain nombre de trames précédentes sur lesquelles le signal de parole ne présente pas d'activité vocale dans la bande i. C'est une fonction des écarts $\left|S_{n-k,i} - \hat{B}_{n-k,i}\right|$ calculés pour un nombre K de trames de silence ($n-k \leq n$). Dans l'exemple représenté, cette fonction est simplement le maximum (bloc 50). Pour chaque trame n, le degré d'activité vocale $\gamma_{n,i}$ est comparé à un seuil (bloc 51)

pour décider si l'écart $\left|S_{n,i} - \hat{B}_{n,i}\right|$, calculé en 52-53, doit ou non être chargé dans une file d'attente 54 de K emplacements organisée en mode premier entré-premier sorti (FIFO). Si $\gamma_{n,i}$ ne dépasse pas le seuil (qui peut être égal à 0 si la fonction g() a la forme de la figure 5), la FIFO 54 n'est pas alimentée, tandis qu'elle l'est dans le cas contraire. La valeur maximale contenue dans la FIFO 54 est alors fournie comme mesure de variabilité $\Delta B_{n,i}^{max}$.

La mesure de variabilité $\Delta B_{n,i}^{max}$ peut, en variante, être obtenue en fonction des valeurs $S_{n,f}$ (et non $S_{n,i}$) et $\hat{B}_{n,i}$. On procède alors de la même manière, sauf que la FIFO 54 contient non pas $\left|S_{n-k,i} - \hat{B}_{n-k,i}\right|$ pour chacune des bandes i, mais plutôt $\max_{f \in [f(i-1), f(i)[} \left|S_{n-k,f} - \hat{B}_{n-k,i}\right|$.

Grâce aux estimations indépendantes des fluctuations à long terme du bruit $\hat{B}_{n,i}$ et de sa variabilité à court terme $\Delta B_{n,i}^{max}$, l'estimateur majoré $\hat{B}'_{n,i}$ procure une excellente robustesse aux bruits musicaux du procédé de débruitage.

Une première phase de la soustraction spectrale est réalisée par le module 55 représenté sur la figure 1. Cette phase fournit, avec la résolution des bandes i ($1 \leq i \leq I$), la réponse en fréquence $H_{n,i}^{1}$ d'un premier filtre de débruitage, en fonction des composantes $S_{n,i}$ et $\hat{B}_{n,i}$ et des coefficients de surestimation $\alpha'_{n,i}$. Ce calcul peut être effectué pour chaque bande i selon la formule :

$$H'_{n,i} = \frac{\max\left\{S_{n,i} - \alpha'_{n,i} \cdot \hat{B}_{n,i} \, , \, \beta'_i \cdot \hat{B}_{n,i}\right\}}{S_{n-\tau4,i}} \qquad (7)$$

où τ4 est un retard entier déterminé tel que τ4≥0 (par exemple τ4=0). Dans l'expression (7), le coefficient β'_i représente, comme le coefficient $\beta_{p,i}$ de la formule (3), un plancher servant classiquement à éviter les valeurs négatives ou trop faibles du signal débruité.

De façon connue (EP-A-0 534 837), le coefficient de surestimation $\alpha'_{n,i}$ pourrait être remplacé dans la formule (7) par un autre coefficient égal à une fonction de $\alpha'_{n,i}$ et d'une estimation du rapport signal-sur-bruit (par exemple $S_{n,i}/\hat{B}_{n,i}$), cette fonction étant décroissante selon la valeur estimée du rapport signal-sur-bruit. Cette fonction est alors égale à $\alpha'_{n,i}$ pour les valeurs les plus faibles du rapport signal-sur-bruit. En effet, lorsque le signal est très bruité, il n'est a priori pas utile de diminuer le facteur de surestimation. Avantageusement, cette fonction décroît vers zéro pour les valeurs les plus élevées du rapport signal/bruit. Ceci permet de protéger les zones les plus énergétiques du spectre, où le signal de parole est le plus significatif, la quantité soustraite du signal tendant alors vers zéro.

Cette stratégie peut être affinée en l'appliquant de manière sélective aux harmoniques de la fréquence tonale (« pitch ») du signal de parole lorsque celui-ci présente une activité vocale.

Ainsi, dans la réalisation représentée sur la figure 1, une seconde phase de débruitage est réalisée par un module 56 de protection des harmoniques. Ce module

calcule, avec la résolution de la transformée de Fourier, la réponse en fréquence $H^2_{n,f}$ d'un second filtre de débruitage en fonction des paramètres $H^1_{n,i}$, $\alpha'_{n,i}$, $\hat{B}_{n,i}$, δ_n, $S_{n,i}$ et de la fréquence tonale $f_p=F_e/T_p$ calculée en dehors des phases de silence par un module d'analyse harmonique 57. En phase de silence ($\delta_n=0$), le module 56 n'est pas en service, c'est-à-dire que $H^2_{n,f} = H^1_{n,i}$ pour chaque fréquence f d'une bande i. Le module 57 peut appliquer toute méthode connue d'analyse du signal de parole de la trame pour déterminer la période T_p, exprimée comme un nombre entier ou fractionnaire d'échantillons, par exemple une méthode de prédiction linéaire.

La protection apportée par le module 56 peut consister à effectuer, pour chaque fréquence f appartenant à une bande i :

$$\begin{cases} H^2_{n,f} = 1 \quad \text{si} \quad \begin{cases} S_{n,i} - \alpha'_{n,i} \cdot \hat{B}_{n,i} > \beta^2_i \cdot \hat{B}_{n,i} & (8) \\ \text{et } \exists \eta \text{ entier } / \left| f - \eta . f_p \right| \leq \Delta f / 2 & (9) \end{cases} \\ H^2_{n,f} = H^1_{n,f} \quad \text{sinon} \end{cases}$$

$\Delta f=F_e/N$ représente la résolution spectrale de la transformée de Fourier. Lorsque $H^2_{n,f}=1$, la quantité soustraite de la composante $S_{n,f}$ sera nulle. Dans ce calcul, les coefficients de plancher β^2_i (par exemple $\beta^2_i = \beta^1_i$) expriment le fait que certaines harmoniques de la fréquence tonale f_p peuvent être masquées par du bruit, de sorte qu'il n'est pas utile de les protéger.

Cette stratégie de protection est de préférence appliquée pour chacune des fréquences les plus proches des

harmoniques de f_p, c'est-à-dire pour η entier quelconque.

Si on désigne par δf_p la résolution fréquentielle avec laquelle le module d'analyse 57 produit la fréquence tonale estimée f_p, c'est-à-dire que la fréquence tonale réelle est comprise entre $f_p - \delta f_p/2$ et $f_p - \delta f_p/2$, alors l'écart entre la η-ième harmonique de la fréquence tonale réelle est son estimation $\eta \times f_p$ (condition (9)) peut aller jusqu'à $\pm \eta \times \delta f_p/2$. Pour les valeurs élevées de η, cet écart peut être supérieur à la demi-résolution spectrale $\Delta f/2$ de la transformée de Fourier. Pour tenir compte de cette incertitude et garantir la bonne protection des harmoniques de la fréquence tonale réelle, on peut protéger chacune des fréquences de l'intervalle $\left[\eta \times f_p - \eta \times \delta f_p/2 \ , \eta \times f_p + \eta \times \delta f_p/2\right]$, c'est-à-dire remplacer la condition (9) ci-dessus par :

$$\exists \eta \text{ entier } / \ \left| f - \eta . f_p \right| \leq \left(\eta . \delta f_p + \Delta f \right)/2 \qquad (9')$$

Cette façon de procéder (condition (9')) présente un intérêt particulier lorsque les valeurs de η peuvent être grandes, notamment dans le cas où le procédé est utilisé dans un système à bande élargie.

Pour chaque fréquence protégée, la réponse en fréquence corrigée $H^2_{n,f}$ peut être égale à 1 comme indiqué ci-dessus, ce qui correspond à la soustraction d'une quantité nulle dans le cadre de la soustraction spectrale, c'est-à-dire à une protection complète de la fréquence en question. Plus généralement, cette réponse en fréquence corrigée $H^2_{n,f}$ pourrait être prise égale à une valeur

comprise entre 1 et $H_{n,f}$ selon le degré de protection souhaité, ce qui correspond à la soustraction d'une quantité inférieure à celle qui serait soustraite si la fréquence en question n'était pas protégée.

Les composantes spectrales $S^2_{n,f}$ d'un signal débruité sont calculées par un multiplieur 58 :

$$S^2_{n,f} = H^2_{n,f} \cdot S_{n,f} \qquad (10)$$

Ce signal $S^2_{n,f}$ est fourni à un module 60 qui calcule, pour chaque trame n, une courbe de masquage en appliquant un modèle psychoacoustique de perception auditive par l'oreille humaine.

Le phénomène de masquage est un principe connu du fonctionnement de l'oreille humaine. Lorsque deux fréquences sont entendues simultanément, il est possible que l'une des deux ne soit plus audible. On dit alors qu'elle est masquée.

Il existe différentes méthodes pour calculer des courbes de masquage. On peut par exemple utiliser celle développée par J.D. Johnston («Transform Coding of Audio Signals Using Perceptual Noise Criteria », IEEE Journal on Selected Area in Communications, Vol. 6, No. 2, février 1988). Dans cette méthode, on travaille dans l'échelle fréquentielle des barks. La courbe de masquage est vue comme la convolution de la fonction d'étalement spectral de la membrane basilaire dans le domaine bark avec le signal excitateur, constitué dans la présente application par le signal $S^2_{n,f}$. La fonction d'étalement spectral peut être modélisée de la manière représentée sur la figure 7. Pour chaque bande de bark, on calcule la contribution des bandes inférieures et supérieures

convoluées par la fonction d'étalement de la membrane basilaire :

$$C_{n,q} = \sum_{q'=0}^{q-1} \frac{S^2_{n,q'}}{\left(10^{-10/10}\right)^{(q-q')}} + \sum_{q'=q+1}^{Q} \frac{S^2_{n,q'}}{\left(10^{25/10}\right)^{(q'-q)}} \qquad (11)$$

où les indices q et q' désignent les bandes de bark ($0 \leq q, q' \leq Q$), et $S^2_{n,q'}$ représente la moyenne des composantes $S^2_{n,f}$ du signal excitateur débruité pour les fréquences discrètes f appartenant à la bande de bark q'.

Le seuil de masquage $M_{n,q}$ est obtenu par le module 60 pour chaque bande de bark q, selon la formule :

$$M_{n,q} = C_{n,q}/R_q \qquad (12)$$

où R_q dépend du caractère plus ou moins voisé du signal. De façon connue, une forme possible de R_q est :

$$10.\log_{10}(R_q) = (A+q).\chi + B.(1-\chi) \qquad (13)$$

avec A=14,5 et B=5,5. χ désigne un degré de voisement du signal de parole, variant entre zéro (pas de voisement) et 1 (signal fortement voisé). Le paramètre χ peut être de la forme connue :

$$\chi = \min\left\{\frac{SFM}{SFM_{max}}, 1\right\} \qquad (12)$$

où SFM représente, en décibels, le rapport entre la moyenne arithmétique et la moyenne géométrique de l'énergie des bandes de bark, et SFM_{max}=-60 dB.

Le système de débruitage comporte encore un module 62 qui corrige la réponse en fréquence du filtre de débruitage, en fonction de la courbe de masquage $M_{n,q}$ calculée par le module 60 et des estimations majorées $\hat{B}'_{n,i}$ calculées par le module 45. Le module 62 décide du niveau

de débruitage qui doit réellement être atteint.

En comparant l'enveloppe de l'estimation majorée du bruit avec l'enveloppe formée par les seuils de masquage $M_{n,q}$, on décide de ne débruiter le signal que dans la mesure où l'estimation majorée $\hat{B}'_{n,i}$ dépasse la courbe de masquage. Ceci évite de supprimer inutilement du bruit masqué par de la parole.

La nouvelle réponse $H^3_{n,f}$, pour une fréquence f appartenant à la bande i définie par le module 12 et à la bande de bark q, dépend ainsi de l'écart relatif entre l'estimation majorée $\hat{B}'_{n,i}$ de la composante spectrale correspondante du bruit et la courbe de masquage $M_{n,q}$, de la manière suivante :

$$H^3_{n,f} = 1 - \left(1 - H^2_{n,f}\right) . \max\left\{\frac{\hat{B}'_{n,i} - M_{n,q}}{\hat{B}'_{n,i}} , 0\right\} \qquad (14)$$

En d'autres termes, la quantité soustraite d'une composante spectrale $S_{n,f}$, dans le processus de soustraction spectrale ayant la réponse fréquentielle $H^3_{n,f}$, est sensiblement égale au minimum entre d'une part la quantité soustraite de cette composante spectrale dans le processus de soustraction spectrale ayant la réponse fréquentielle $H^2_{n,f}$, et d'autre part la fraction de l'estimation majorée $\hat{B}'_{n,i}$ de la composante spectrale correspondante du bruit qui, le cas échéant, dépasse la courbe de masquage $M_{n,q}$.

La figure 8 illustre le principe de la correction appliquée par le module 62. Elle montre schématiquement un exemple de courbe de masquage $M_{n,q}$ calculée sur la base

des composantes spectrales $S_{n,f}$ du signal débruité, ainsi que l'estimation majorée $\hat{B}'_{n,i}$ du spectre du bruit. La quantité finalement soustraite des composantes $S_{n,f}$ sera celle représentée par les zones hachurées, c'est-à-dire limitée à la fraction de l'estimation majorée $\hat{B}'_{n,i}$ des composantes spectrales du bruit qui dépasse la courbe de masquage.

Cette soustraction est effectuée en multipliant la réponse fréquentielle $H^3_{n,f}$ du filtre de débruitage par les composantes spectrales $S_{n,f}$ du signal de parole (multiplieur 64). Un module 65 reconstruit alors le signal débruité dans le domaine temporel, en opérant la transformée de Fourier rapide inverse (TFRI) inverse des échantillons de fréquence $S^3_{n,f}$ délivrés par le multiplieur 64. Pour chaque trame, seuls les N/2=128 premiers échantillons du signal produit par le module 65 sont délivrés comme signal débruité final s^3, après reconstruction par addition-recouvrement avec les N/2=128 derniers échantillons de la trame précédente (module 66).

La figure 9 montre une forme de réalisation préférée d'un système de débruitage mettant en œuvre l'invention. Ce système comporte un certain nombre d'éléments semblables à des éléments correspondants du système de la figure 1, pour lesquels on a utilisé les mêmes références numériques. Ainsi, les modules 10, 11, 12, 15, 16, 45 et 55 fournissent notamment les quantités $S_{n,i}$, $\hat{B}_{n,i}$, $\alpha'_{n,i}$, $\hat{B}'_{n,i}$ et $H^3_{n,f}$ pour effectuer le débruitage sélectif.

La résolution en fréquence de la transformée de

Fourier rapide 11 est une limitation du système de la figure 1. En effet, la fréquence faisant l'objet de la protection par le module 56 n'est pas nécessairement la fréquence tonale précise f_p, mais la fréquence la plus proche de celle-ci dans le spectre discret. Dans certains cas, on peut alors protéger des harmoniques relativement éloignées de celle de la fréquence tonale. Le système de la figure 9 pallie cet inconvénient grâce à un conditionnement approprié du signal de parole.

Dans ce conditionnement, on modifie la fréquence d'échantillonnage du signal de telle sorte que la période $1/f_p$ couvre exactement un nombre entier de temps d'échantillon du signal conditionné.

De nombreuses méthodes d'analyse harmonique pouvant être mises en œuvre par le module 57 sont capables de fournir une valeur fractionnaire du retard T_p, exprimé en nombre d'échantillons à la fréquence d'échantillonnage initiale F_e. On choisit alors une nouvelle fréquence d'échantillonnage f_e de telle sorte qu'elle soit égale à un multiple entier de la fréquence tonale estimée, soit $f_e=p.f_p=p.F_e/T_p=K.F_e$, avec p entier. Afin de ne pas perdre d'échantillons de signal, il convient que f_e soit supérieure à F_e. On peut notamment imposer qu'elle soit comprise entre F_e et $2F_e$ ($1 \leq K \leq 2$), pour faciliter la mise en œuvre du conditionnement.

Bien entendu, si aucune activité vocale n'est détectée sur la trame courante ($\delta_n \neq 0$) , ou si le retard T_p estimé par le module 57 est entier, il n'est pas nécessaire de conditionner le signal.

Afin que chacune des harmoniques de la fréquence tonale corresponde également à un nombre entier

d'échantillons du signal conditionné, l'entier p doit être un diviseur de la taille N de la fenêtre de signal produite par le module 10 : $N=\alpha p$, avec α entier. Cette taille N est usuellement une puissance de 2 pour la mise en œuvre de la TFR. Elle est de 256 dans l'exemple considéré.

La résolution spectrale Δf de la transformée de Fourier discrète du signal conditionné est donnée par $\Delta f=p.f_p/N=f_p/\alpha$. On a donc intérêt à choisir p petit de façon à maximiser α, mais suffisamment grand pour suréchantillonner. Dans l'exemple considéré, où F_e=8 kHz et N=256, les valeurs choisies pour les paramètres p et α sont indiquées dans le tableau I.

500 Hz < f_p < 1000 Hz	8 < T_p < 16	p = 16	α = 16
250 Hz < f_p < 500 Hz	16 < T_p < 32	p = 32	α = 8
125 Hz < f_p < 250 Hz	32 < T_p < 64	p = 64	α = 4
62,5 Hz < f_p < 125 Hz	64 < T_p < 128	p = 128	α = 2
31,25 Hz < f_p < 62,5 Hz	128 < T_p < 256	p = 256	α = 1

Tableau I

Ce choix est effectué par un module 70 selon la valeur du retard T_p fournie par le module d'analyse harmonique 57. Le module 70 fournit le rapport K entre les fréquences d'échantillonnage à trois modules de changement de fréquence 71, 72, 73.

Le module 71 sert à transformer les valeurs $S_{n,i}$, $\hat{B}_{n,i}$, $\alpha'_{n,i}$, $\hat{B}'_{n,i}$ et $H'_{n,f}$, relatives aux bandes i définies par le module 12, dans l'échelle des fréquences modifiées (fréquence d'échantillonnage f_e). Cette transformation

consiste simplement à dilater les bandes i dans le facteur K. Les valeurs ainsi transformées sont fournies au module 56 de protection des harmoniques.

Celui-ci opère alors de la même manière que précédemment pour fournir la réponse en fréquence $H^2_{n,f}$ du filtre de débruitage. Cette réponse $H^2_{n,f}$ est obtenue de la même manière que dans le cas de la figure 1 (conditions (8) et (9)), à cette différence près que, dans la condition (9), la fréquence tonale $f_p=f_e/p$ est définie selon la valeur du retard entier p fourni par le module 70, la résolution en fréquence Δf étant également fournie par ce module 70.

Le module 72 procède au suréchantillonnage de la trame de N échantillons fournie par le module de fenêtrage 10. Le suréchantillonnage dans un facteur K rationnel (K=K1/K2) consiste à effectuer d'abord un suréchantillonnage dans le facteur entier K1, puis un sous-échantillonnage dans le facteur entier K2. Ces suréchantillonnage et sous-échantillonnage dans des facteurs entiers peuvent être effectués classiquement au moyen de bancs de filtres polyphase.

La trame de signal conditionné s' fournie par le module 72 comporte KN échantillons à la fréquence f_e. Ces échantillons sont adressés à un module 75 qui calcule leur transformée de Fourier. La transformation peut être effectuée à partir de deux blocs de N=256 échantillons : l'un constitué par les N premiers échantillons de la trame de longueur KN du signal conditionné s', et l'autre par les N derniers échantillons de cette trame. Les deux blocs présentent donc un recouvrement de (2-K)×100%. Pour chacun des deux blocs, on obtient un jeu de composantes de

Fourier $S_{n,f}$. Ces composantes $S_{n,f}$ sont fournies au multiplieur 58, qui les multiplie par la réponse spectrale $H^2_{n,f}$ pour délivrer les composantes spectrales $S^2_{n,f}$ du premier signal débruité.

Ces composantes $S^2_{n,f}$ sont adressées au module 60 qui calcule les courbes de masquage de la manière précédemment indiquée.

De préférence, dans ce calcul des courbes de masquage, la grandeur χ désignant le degré de voisement du signal de parole (formule (13)) est prise de la forme $\chi=1-H$, où H est une entropie de l'autocorrelation des composantes spectrales $S^2_{n,f}$ du signal conditionné débruité. Les autocorrelations A(k) sont calculées par un module 76, par exemple selon la formule :

$$A(k) = \frac{\sum_{f=0}^{N/2-1} S^2_{n,f} \cdot S^2_{n,f+k}}{\sum_{f=0}^{N/2-1} \sum_{f'=0}^{N/2-1} S^2_{n,f} \cdot S^2_{n,f+f'}} \qquad (15)$$

Un module 77 calcule ensuite l'entropie normalisée H, et la fournit au module 60 pour le calcul de la courbe de masquage (voir S.A. McClellan et al : « Spectral Entropy : an Alternative Indicator for Rate Allocation ? », Proc. ICASSP'94, pages 201-204) :

$$H = \frac{\sum_{k=0}^{N/2-1} A(k) \cdot \log[A(k)]}{\log N/2)} \qquad (16)$$

Grâce au conditionnement du signal, ainsi qu'à son débruitage par le filtre $H^2_{n,f}$, l'entropie normalisée H

constitue une mesure de voisement très robuste au bruit et aux variations de la fréquence tonale.

Le module de correction 62 opère de la même manière que celui du système de la figure 1, en tenant compte du bruit surestimé $\hat{B}'_{n,i}$ remis à l'échelle par le module de changement de fréquence 71. Il fournit la réponse en fréquence $H^3_{n,f}$ du filtre de débruitage définitif, qui est multipliée par les composantes spectrales $S_{n,f}$ du signal conditionné par le multiplieur 64. Les composantes $S^3_{n,f}$ qui en résultent sont ramenées dans le domaine temporel par le module de TFRI 65. En sortie de cette TFRI 65, un module 80 combine, pour chaque trame, les deux blocs de signal issus du traitement des deux blocs recouvrants délivrés par la TFR 75. Cette combinaison peut consister en une somme avec pondération de Hamming des échantillons, pour former une trame de signal conditionné débruité de KN échantillons.

Le signal conditionné débruité fourni par le module 80 fait l'objet d'un changement de fréquence d'échantillonnage par le module 73. Sa fréquence d'échantillonnage est ramenée à $F_e=f_e/K$ par les opérations inverses de celles effectuées par le module 75. Le module 73 délivre N=256 échantillons par trame. Après la reconstruction par addition-recouvrement avec les N/2=128 derniers échantillons de la trame précédente, seuls les N/2=128 premiers échantillons de la trame courante sont finalement conservés pour former le signal débruité final s^3 (module 66).

Dans une forme de réalisation préférée, un module 82 gère les fenêtres formées par le module 10 et sauvegardées par le module 66, de façon telle qu'on

sauvegarde un nombre M d'échantillons égal à un multiple entier de $T_p=F_e/f_p$. On évite ainsi les problèmes de discontinuité de phase entre les trames. De façon correspondante, le module de gestion 82 commande le module de fenêtrage 10 pour que le recouvrement entre la trame courante et la prochaine corresponde à N-M. Il sera tenu de ce recouvrement de N-M échantillons dans la somme à recouvrement effectuée par le module 66 lors du traitement de la prochaine trame. A partir de la valeur de T_p fournie par le module d'analyse harmonique 57, le module 82 calcule le nombre d'échantillons à sauvegarder $M=T_p \times E[N/(2T_p)]$, E[] désignant la partie entière, et commande de façon correspondante les modules 10 et 66.

Dans le mode de réalisation qu'on vient de décrire, la fréquence tonale est estimée de façon moyenne sur la trame. Or la fréquence tonale peut varier quelque peu sur cette durée. Il est possible de tenir compte de ces variations dans le cadre de la présente invention, en conditionnant le signal de façon à obtenir artificiellement une fréquence tonale constante dans la trame.

Pour cela, on a besoin que le module 57 d'analyse harmonique fournisse les intervalles de temps entre les ruptures consécutives du signal de parole attribuables à des fermetures de la glotte du locuteur intervenant pendant la durée de la trame. Des méthodes utilisables pour détecter de telles micro-ruptures sont bien connues dans le domaine de l'analyse harmonique des signaux de paroles. On pourra à cet égard consulter les articles suivants : M. BASSEVILLE et al., « Sequential detection of abrupt changes in spectral characteristics of digital signals », IEEE Trans. on Information Theory, 1983, Vol. IT-29, n°5, pages 708-723. ; R. ANDRE-OBRECHT, « A new

statistical approach for the automatic segmentation of continuous speech signals », IEEE Trans. on Acous., Speech and Sig. Proc., Vol. 36, N°1, janvier 1988 ; et C. MURGIA et al., « An algorithm for the estimation of glottal closure instants using the sequential detection of abrupt changes in speech signals », Signal Processing VII, 1994, pages 1685-1688.

Le principe de ces méthodes est d'effectuer un test statistique entre deux modèles, l'un à court terme et l'autre à long terme. Les deux modèles sont des modèles adaptatifs de prédiction linéaire. La valeur de ce test statistique w_m est la somme cumulée du rapport de vraisemblance a posteriori de deux distributions, corrigée par la divergence de Kullback. Pour une distribution de résidus ayant une statistique gaussienne, cette valeur w_m est donnée par :

$$w_m = \frac{1}{2}\left[\frac{2\, e_m^0 . e_m^1}{\sigma_1^2} - \left(1 + \frac{\sigma_0^2}{\sigma_1^2}\right).\frac{\left(e_m^0\right)^2}{\sigma_0^2} + \left(1 - \frac{\sigma_0^2}{\sigma_1^2}\right)\right] \qquad (17)$$

où e_m^0 et σ_0^2 représentent le résidu calculé au moment de l'échantillon m de la trame et la variance du modèle à long terme, e_m^1 et σ_1^2 représentant de même le résidu et la variance du modèle à court terme. Plus les deux modèles sont proches, plus la valeur w_m du test statistique est proche de 0. Par contre, lorsque les deux modèles sont éloignés l'un de l'autre, cette valeur w_m devient négative, ce qui dénote une rupture R du signal.

La figure 10 montre ainsi un exemple possible d'évolution de la valeur w_m, montrant les ruptures R du signal de parole. Les intervalles de temps t_r (r = 1,2,...) entre deux ruptures consécutives R sont

calculés, et exprimés en nombre d'échantillons du signal de parole. Chacun de ces intervalles t_r est inversement proportionnel à la fréquence tonale f_p, qui est ainsi estimée localement : $f_p=F_e/t_r$ sur le r-ième intervalle.

On peut alors corriger les variations temporelles de la fréquence tonale (c'est-à-dire le fait que les intervalles t_r ne sont pas tous égaux sur une trame donnée), afin d'avoir une fréquence tonale constante dans chacune des trames d'analyse. Cette correction est effectuée par une modification de la fréquence d'échantillonnage sur chaque intervalle t_r, de façon à obtenir, après suréchantillonnage, des intervalles constants entre deux ruptures glottiques. On modifie donc la durée entre deux ruptures en faisant un suréchantillonnage dans un rapport variable, de façon à se caler sur l'intervalle le plus grand. De plus, on fait en sorte de respecter la contrainte de conditionnement selon laquelle la fréquence de suréchantillonnage est multiple de la fréquence tonale estimée.

La figure 11 montre les moyens utilisés pour calculer le conditionnement du signal dans ce dernier cas. Le module 57 d'analyse harmonique est réalisé de façon à mettre en œuvre la méthode d'analyse ci-dessus, et à fournir les intervalles t_r relatifs à la trame de signal produite par le module 10. Pour chacun de ces intervalles, le module 70 (bloc 90 sur la figure 11) calcule le rapport de suréchantillonnage $K_r=p_r/t_r$, où l'entier p_r est donné par la troisième colonne du tableau I lorsque t_r prend les valeurs indiquées dans la deuxième colonne. Ces rapports de suréchantillonnage K_r sont fournis aux modules de changement de fréquence 72 et 73, pour que les interpolations soient effectuées avec le rapport

d'échantillonnage K_r sur l'intervalle de temps correspondant t_r.

Le plus grand T_p des intervalles de temps t_r fournis par le module 57 pour une trame est sélectionné par le module 70 (bloc 91 sur la figure 11) pour obtenir un couple p,α comme indiqué dans le tableau I. La fréquence d'échantillonnage modifiée est alors $f_e=p.F_e/T_p$ comme précédemment, la résolution spectrale Δf de la transformée de Fourier discrète du signal conditionné étant toujours donnée par $\Delta f=F_e/(\alpha.T_p)$. Pour le module de changement de fréquence 71, le rapport de suréchantillonnage K est donné par $K=p/T_p$ (bloc 92). Le module 56 de protection des harmoniques de la fréquence tonale opère de la même manière que précédemment, en utilisant pour la condition (9) la résolution spectrale Δf fournie par le bloc 91 et la fréquence tonale $f_p=f_e/p$ définie selon la valeur du retard entier p fournie par le bloc 91.

Cette forme de réalisation de l'invention implique également une adaptation du module 82 de gestion des fenêtres. Le nombre M d'échantillons du signal débruité à sauvegarder sur la trame courante correspond ici à un nombre entier d'intervalles de temps t_r consécutifs entre deux ruptures glottiques (voir figure 10). Cette disposition évite les problèmes de discontinuité de phase entre trames, tout en tenant compte des variations possibles des intervalles de temps t_r sur une trame.

REVENDICATIONS

1. Procédé de débruitage d'un signal de parole numérique (s) traité par trames successives, dans lequel :

- on calcule des composantes spectrales ($S_{n,f}$, $S_{n,i}$) du signal de parole sur chaque trame ;

- on calcule pour chaque trame des estimations majorées ($\hat{B}'_{n,i}$) de composantes spectrales du bruit compris dans le signal de parole ;

- on effectue une soustraction spectrale comportant au moins une première étape de soustraction dans laquelle on soustrait respectivement, de chaque composante spectrale ($S_{n,f}$) du signal de parole sur la trame, une première quantité dépendant de paramètres incluant l'estimation majorée ($\hat{B}'_{n,i}$) de la composante spectrale correspondante du bruit pour ladite trame, de manière à obtenir des composantes spectrales ($S^2_{n,f}$) d'un premier signal débruité ; et

- on applique au résultat de la soustraction spectrale une transformation vers le domaine temporel pour construire un signal de parole débruité (s^3),

caractérisé en ce que la soustraction spectrale comporte en outre les étapes suivantes :

- le calcul d'une courbe de masquage ($M_{n,q}$) en appliquant un modèle de perception auditive à partir des composantes spectrales ($S^2_{n,f}$) du premier signal débruité ;

- la comparaison des estimations majorées ($\hat{B}'_{n,i}$) des composantes spectrales du bruit pour la trame à la courbe de masquage calculée ($M_{n,q}$) ; et

- une seconde étape de soustraction dans laquelle on soustrait respectivement, de chaque composante

spectrale ($S_{n,f}$) du signal de parole sur la trame, une seconde quantité dépendant de paramètres incluant un écart entre l'estimation majorée de la composante spectrale correspondante du bruit et la courbe de masquage calculée.

2. Procédé selon la revendication 1, dans lequel ladite seconde quantité relative à une composante spectrale ($S_{n,f}$) du signal de parole sur la trame est sensiblement égale au minimum entre la première quantité correspondante et la fraction de l'estimation majorée ($\hat{B}'_{n,i}$) de la composante spectrale correspondante du bruit qui dépasse la courbe de masquage ($M_{n,q}$).

3. Procédé selon la revendication 1 ou 2, dans lequel on effectue une analyse harmonique du signal de parole pour estimer une fréquence tonale (f_p) du signal de parole sur chaque trame où il présente une activité vocale.

4. Procédé selon la revendication 3, dans lequel les paramètres dont dépendent les premières quantités soustraites incluent la fréquence tonale estimée (f_p).

5. Procédé selon la revendication 4, dans lequel la première quantité soustraite d'une composante spectrale donnée ($S_{n,f}$) du signal de parole est plus faible si ladite composante spectrale correspond à la fréquence la plus proche d'un multiple entier de la fréquence tonale estimée (f_p) que si ladite composante spectrale ne correspond pas à la fréquence la plus proche d'un multiple entier de la fréquence tonale estimée.

6. Procédé selon la revendication 4 ou 5, dans lequel les quantités respectivement soustraites des composantes spectrales ($S_{n,f}$) du signal de parole correspondant aux fréquences les plus proches des multiples entiers de la fréquence tonale estimée (f_p) sont sensiblement nulles.

7. Procédé selon l'une quelconque des revendications 3 à 6, dans lequel, après avoir estimé la fréquence tonale (f_p) du signal de parole sur une trame, on conditionne le signal de parole de la trame en le suréchantillonnant à une fréquence de suréchantillonnage (f_e) multiple de la fréquence tonale estimée, et on calcule les composantes spectrales ($S_{n,f}$) du signal de parole sur la trame sur la base du signal conditionné (s') pour leur soustraire lesdites quantités.

8. Procédé selon la revendication 7, dans lequel on calcule des composantes spectrales ($S_{n,f}$) du signal de parole en distribuant le signal conditionné (s') par blocs de N échantillons soumis à une transformation dans le domaine fréquentiel, et dans lequel le rapport (p) entre la fréquence de suréchantillonnage (f_e) et la fréquence tonale estimée est un diviseur du nombre N.

9. Procédé selon la revendication 7 ou 8, dans lequel on estime un degré de voisement (χ) du signal de parole sur la trame à partir d'un calcul de l'entropie (H) de l'autocorrelation des composantes spectrales calculées sur la base du signal conditionné.

10. Procédé selon la revendication 9, dans lequel lesdites composantes spectrales ($S^2_{n,f}$) dont on calcule l'autocorrelation (H) sont celles calculées sur la base du signal conditionné (s') après soustraction desdites premières quantités.

11. Procédé selon la revendication 9 ou 10, dans lequel le degré de voisement (χ) est mesuré à partir une entropie normalisée H de la forme :

$$H = \frac{\sum_{k=0}^{N/2-1} A(k) \cdot \log[A(k)]}{\log(N/2)}$$

où N est le nombre d'échantillons utilisés pour calculer les composantes spectrales ($S_{n,f}$) sur la base du signal conditionné (s'), et A(k) est l'autocorrelation normalisée définie par :

$$A(k) = \frac{\sum_{f=0}^{N/2-1} S_{n,f}^2 \cdot S_{n,f+k}^2}{\sum_{f=0}^{N/2-1} \sum_{f'=0}^{N/2-1} S_{n,f}^2 \cdot S_{n,f+f'}^2}$$

$S_{n,f}^2$ désignant la composante spectrale de rang f calculée sur la base du signal conditionné.

12. Procédé selon la revendication 11, dans lequel le calcul de la courbe de masquage ($M_{n,q}$) fait intervenir le degré de voisement (χ) mesuré par l'entropie normalisée H.

13. Procédé selon l'une quelconque des revendications 3 à 12, dans lequel, après le traitement de chaque trame, on conserve, parmi les échantillons du signal de parole débruité fournis par ce traitement, un nombre d'échantillons (M) égal à un multiple entier de fois le rapport (T_p) entre la fréquence d'échantillonnage (F_e) et la fréquence tonale estimée (f_p).

14. Procédé selon l'une quelconque des revendications 3 à 12, dans lequel l'estimation de la fréquence tonale du signal de parole sur une trame comporte les étapes suivantes :

- on estime des intervalles de temps (t_r) entre deux ruptures consécutives (R) du signal attribuables à des fermetures de la glotte du locuteur intervenant pendant la durée de la trame, la fréquence tonale estimée étant inversement proportionnelle auxdits intervalles de temps ;

- on interpole le signal de parole dans lesdits intervalles de temps, afin que le signal conditionné (s') résultant de cette interpolation présente un intervalle de temps constant entre deux ruptures consécutives.

15. Procédé selon la revendication 14, dans lequel, après le traitement de chaque trame, on conserve, parmi les échantillons du signal de parole débruité fournis par ce traitement, un nombre d'échantillons (M) correspondant à un nombre entier d'intervalles de temps estimés (t_r).

16. Procédé selon l'une quelconque des revendications précédentes, dans lequel on estime dans le domaine spectral des valeurs d'un rapport signal-sur-bruit que présente le signal de parole (s) sur chaque trame, et dans lequel les paramètres dont dépendent les premières quantités soustraites incluent les valeurs estimées du rapport signal-sur-bruit, la première quantité soustraite de chaque composante spectrale ($S_{n,f}$) du signal de parole sur la trame étant une fonction décroissante de la valeur estimée correspondante du rapport signal-sur-bruit.

17. Procédé selon la revendication 16, dans lequel ladite fonction décroît vers zéro pour les valeurs les plus élevées du rapport signal-sur-bruit.

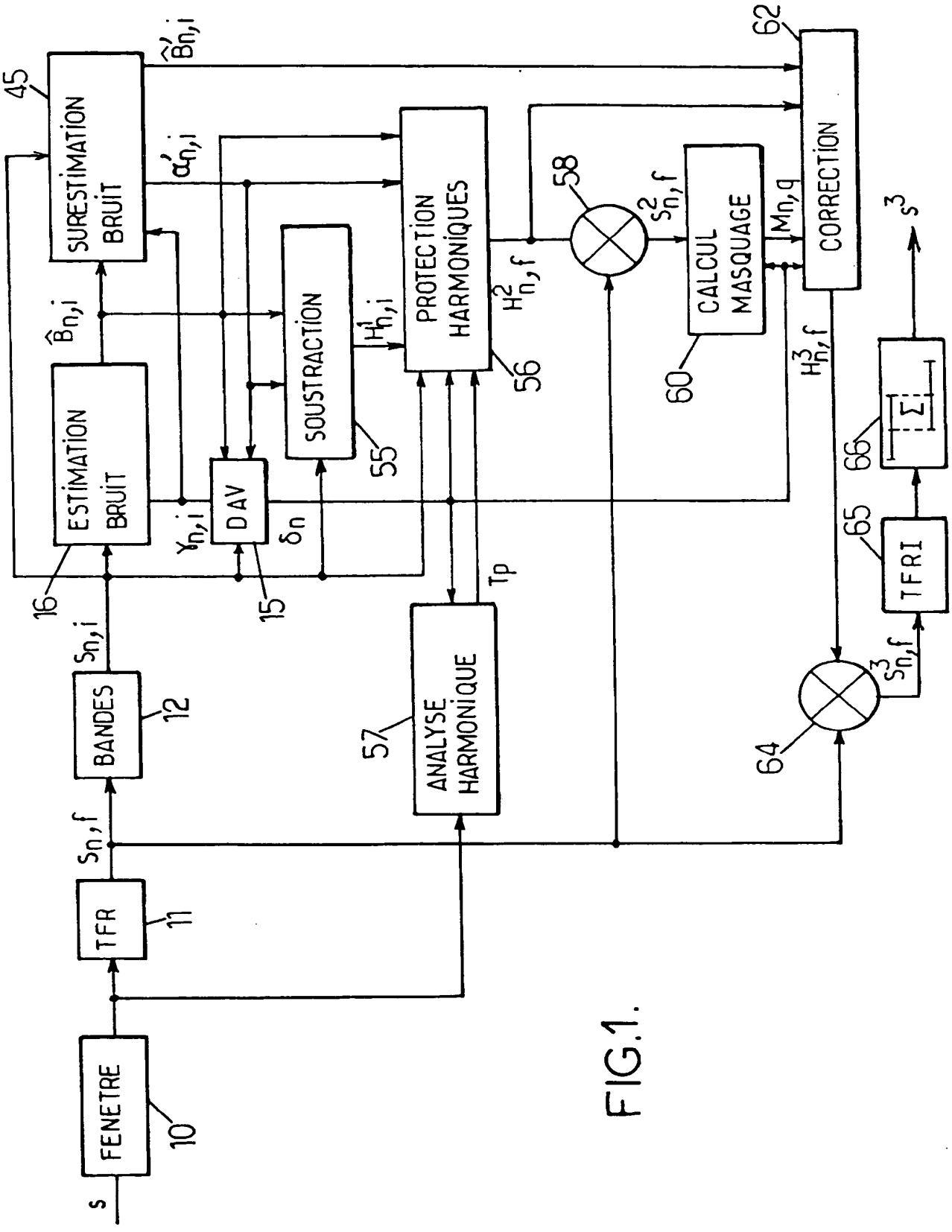

s
FENETRE
10
TFR
11
BANDES
12
ESTIMATION BRUIT
16
SURESTIMATION BRUIT
45
DAV
15
SOUSTRACTION
55
PROTECTION HARMONIQUES
56
ANALYSE HARMONIQUE
57
CALCUL MASQUAGE
60
CORRECTION
62
TFRI
65
64
66
Σ
$\hat{B}_{n,i}$
$\hat{B}'_{n,i}$
$\alpha'_{n,i}$
$\gamma_{n,i}$
δ_n
$S_{n,i}$
$S_{n,f}$
$H^1_{n,i}$
$H^2_{n,f}$
$H^3_{n,f}$
$S^2_{n,f}$
$S^3_{n,f}$
$M_{n,q}$
Tp
s^3

FIG.1.

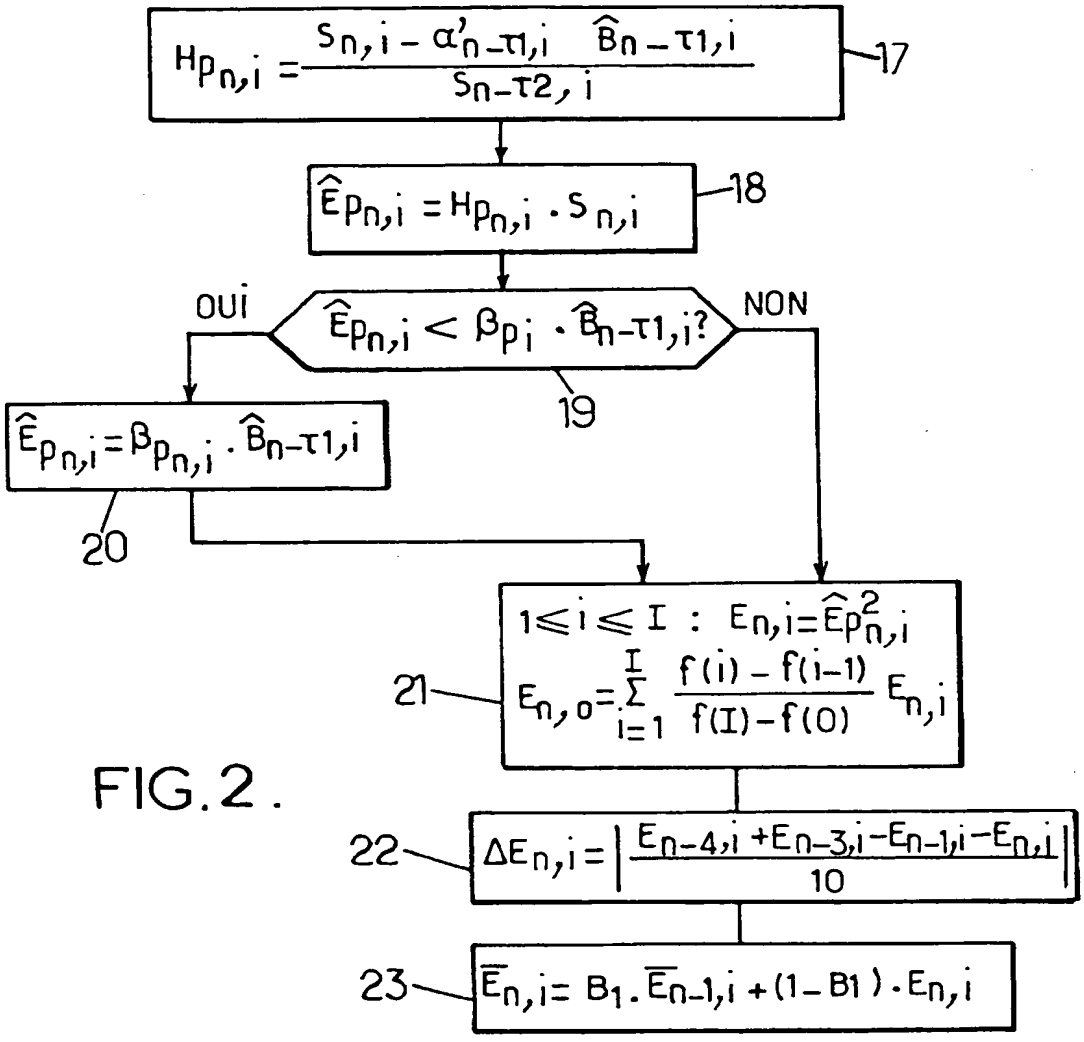

$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau1,i} \ \hat{B}_{n-\tau1,i}}{S_{n-\tau2,i}}$
17
$\hat{E}p_{n,i} = Hp_{n,i} \cdot S_{n,i}$
18
OUI
$\hat{E}p_{n,i} < \beta p_i \cdot \hat{B}_{n-\tau1,i}$?
NON
19
$\hat{E}p_{n,i} = \beta p_{n,i} \cdot \hat{B}_{n-\tau1,i}$
20
$1 \leqslant i \leqslant I : E_{n,i} = \hat{E}p^2_{n,i}$
$E_{n,0} = \sum_{i=1}^{I} \frac{f(i) - f(i-1)}{f(I) - f(0)} E_{n,i}$
21
$\Delta E_{n,i} = \left| \frac{E_{n-4,i} + E_{n-3,i} - E_{n-1,i} - E_{n,i}}{10} \right|$
22
$\overline{E}_{n,i} = B_1 \cdot \overline{E}_{n-1,i} + (1 - B1) \cdot E_{n,i}$
23

FIG.2.

FIG.4.

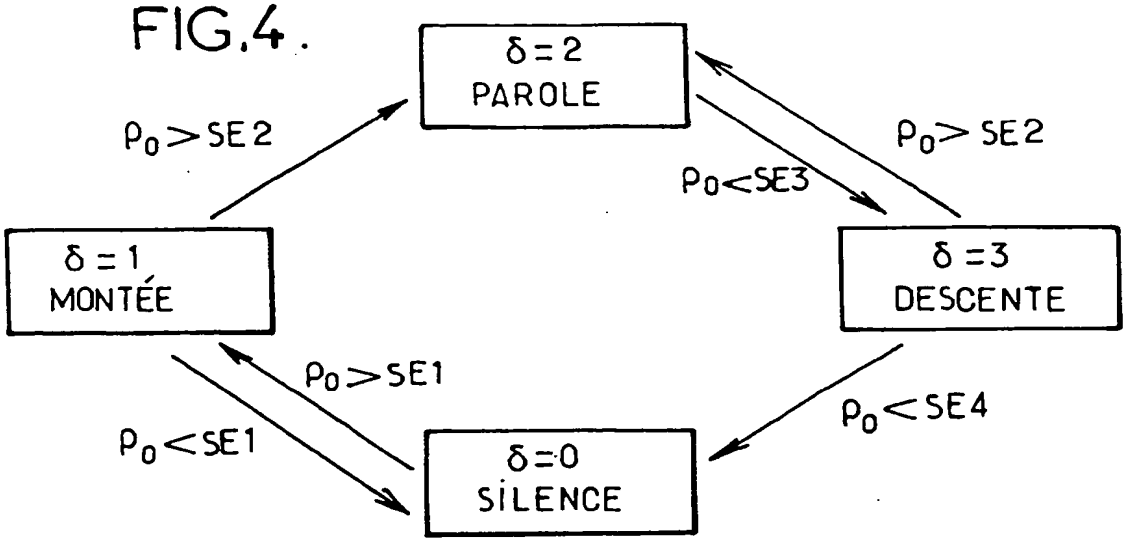
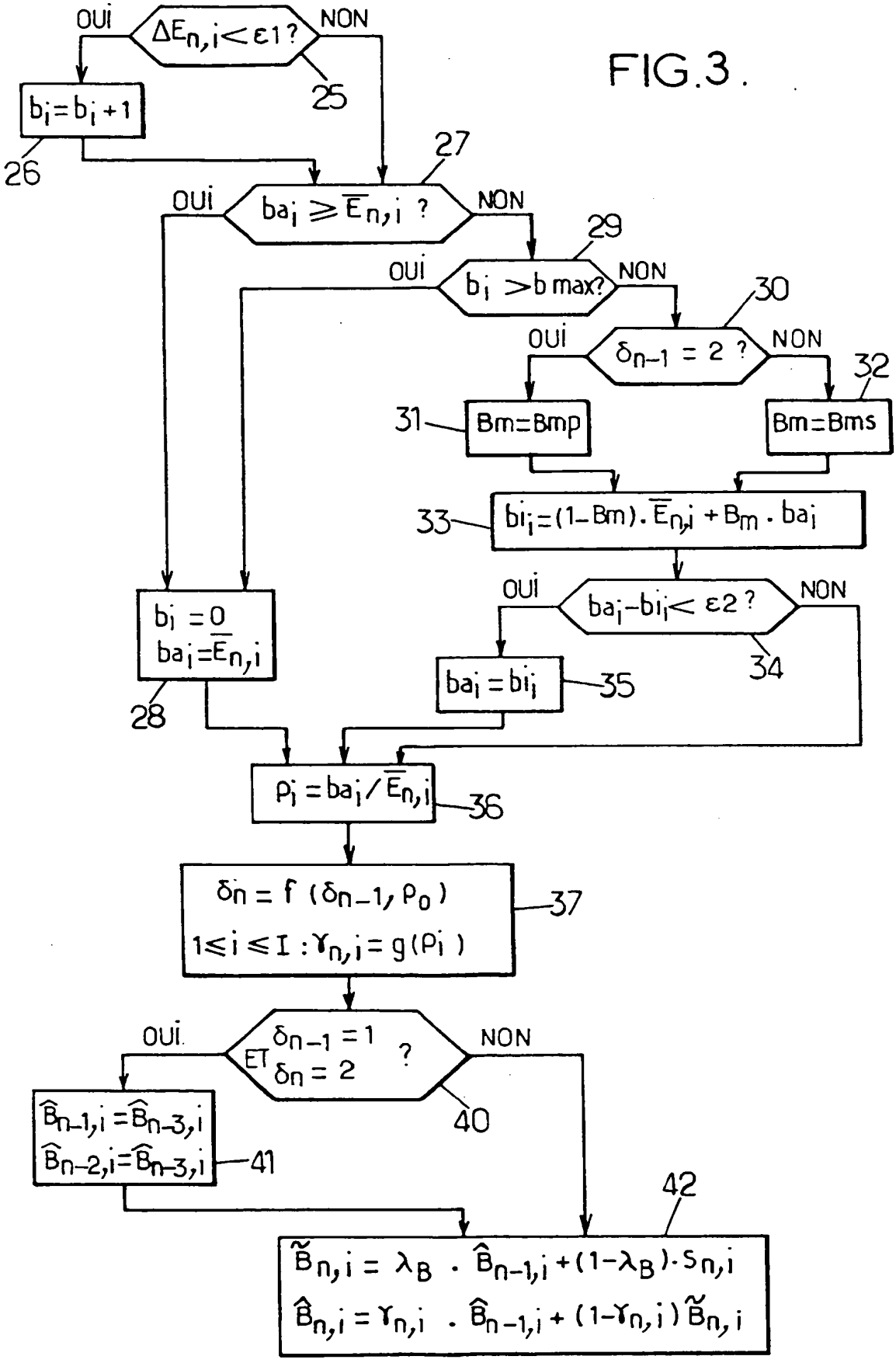

δ = 2
PAROLE
$p_0 > SE2$
$p_0 < SE3$
$p_0 > SE2$
δ = 1
MONTÉE
δ = 3
DESCENTE
$p_0 > SE1$
$p_0 < SE1$
$p_0 < SE4$
δ = 0
SILENCE

FIG.3.


OUI
$\Delta E_{n,i} < \varepsilon 1$?
NON
25
$b_i = b_i + 1$
26
27
OUI
$ba_i \geq \overline{E}_{n,i}$?
NON
29
OUI
$b_i > b\,max$?
NON
30
OUI
$\delta_{n-1} = 2$?
NON
32
31
Bm=Bmp
Bm=Bms
33
$bi_i = (1-Bm) \cdot \overline{E}_{n,i} + B_m \cdot ba_i$
OUI
$ba_i - bi_i < \varepsilon 2$?
NON
34
$b_i = 0$
$ba_i = \overline{E}_{n,i}$
28
$ba_i = bi_i$
35
$\rho_i = ba_i / \overline{E}_{n,i}$
36
$\delta_n = f(\delta_{n-1}, \rho_0)$
$1 \leq i \leq I : \gamma_{n,i} = g(\rho_i)$
37
OUI
$\delta_{n-1} = 1$ ET $\delta_n = 2$?
NON
40
$\hat{B}_{n-1,i} = \hat{B}_{n-3,i}$
$\hat{B}_{n-2,i} = \hat{B}_{n-3,i}$
41
42
$\tilde{B}_{n,i} = \lambda_B \cdot \hat{B}_{n-1,i} + (1-\lambda_B) \cdot s_{n,i}$
$\hat{B}_{n,i} = \gamma_{n,i} \cdot \hat{B}_{n-1,i} + (1-\gamma_{n,i}) \tilde{B}_{n,i}$

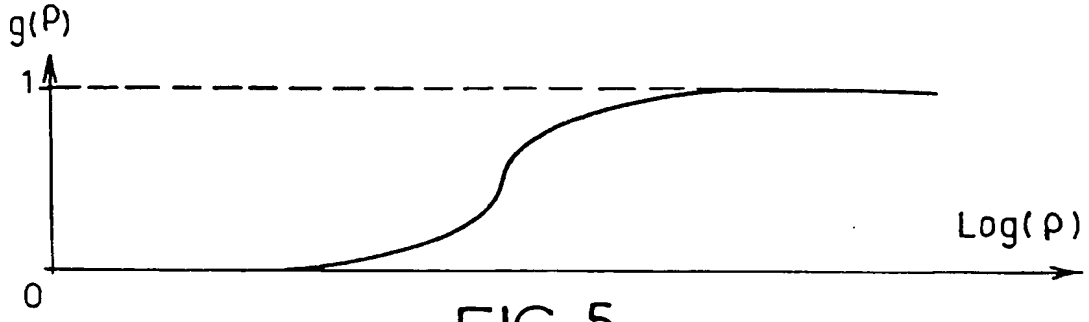

g(ρ)
1
0
Log(ρ)

FIG. 5.

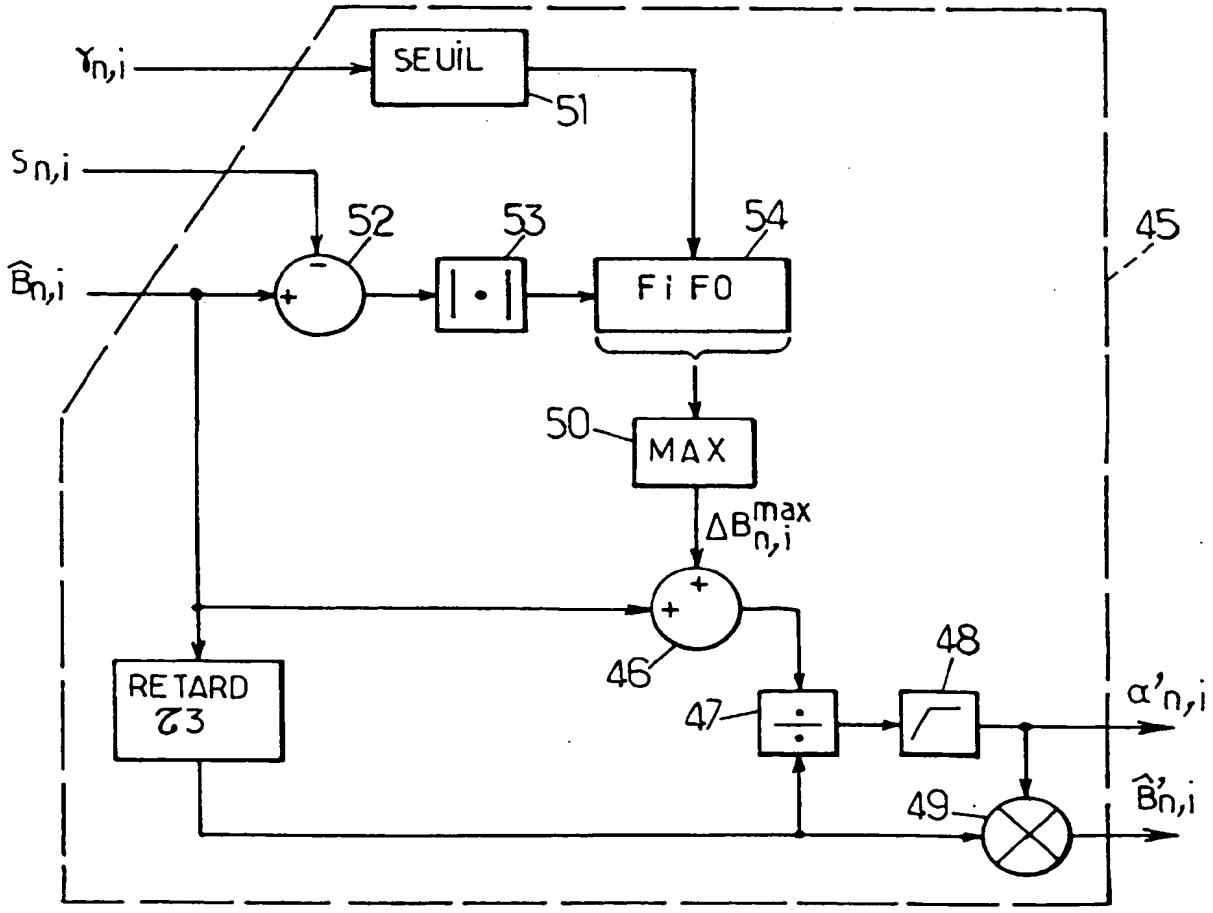

$\gamma_{n,i}$
SEUIL
51
$s_{n,i}$
52
53
54
45
$\hat{B}_{n,i}$
FIFO
50
MAX
$\Delta B^{max}_{n,i}$
46
48
47
$\alpha'_{n,i}$
RETARD
23
49
$\hat{B}'_{n,i}$

FIG. 6.

FIG. 7.

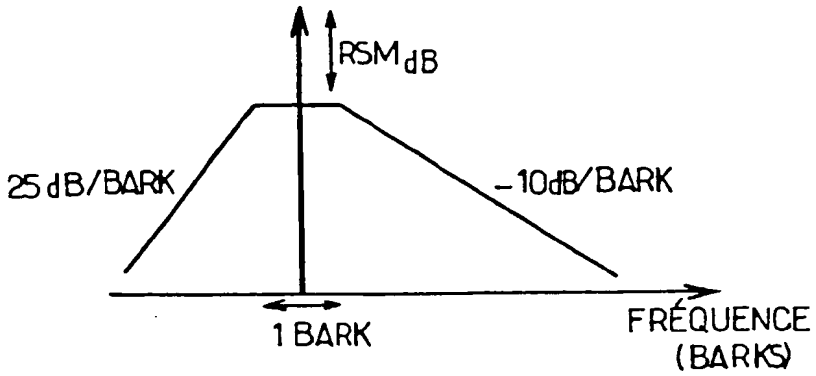
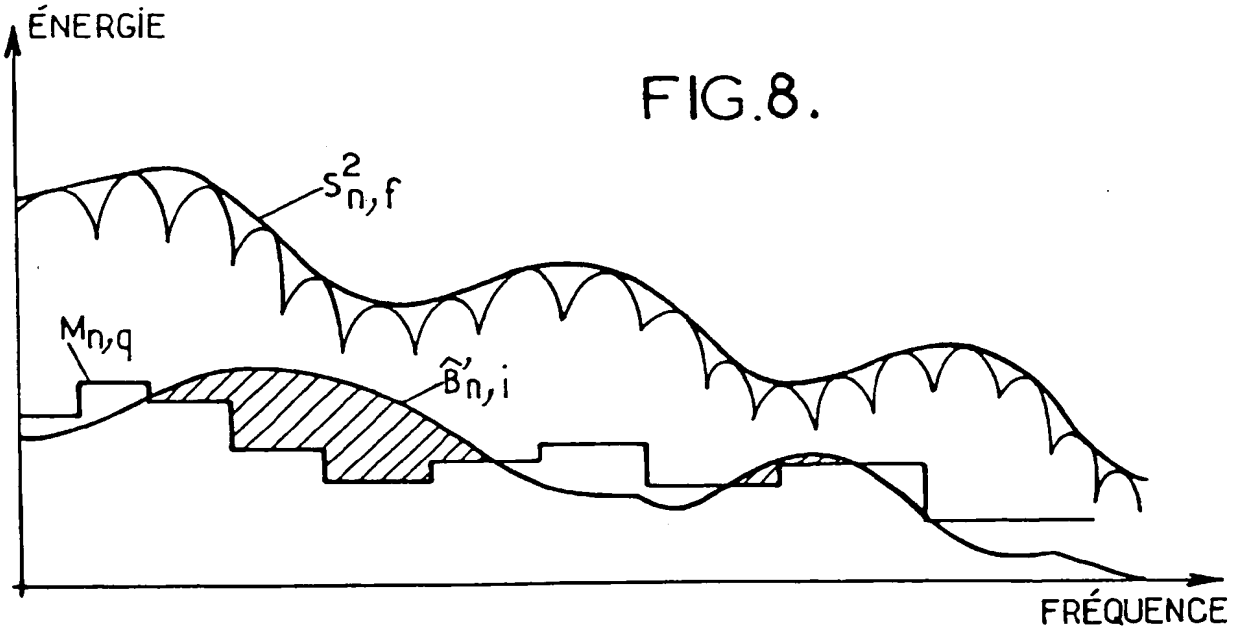

RSM dB
25 dB/BARK
-10dB/BARK
1 BARK
FRÉQUENCE
(BARKS)

FIG. 8.


ÉNERGIE
$s^2_{n,f}$
$M_{n,q}$
$\hat{B}'_{n,i}$
FRÉQUENCE

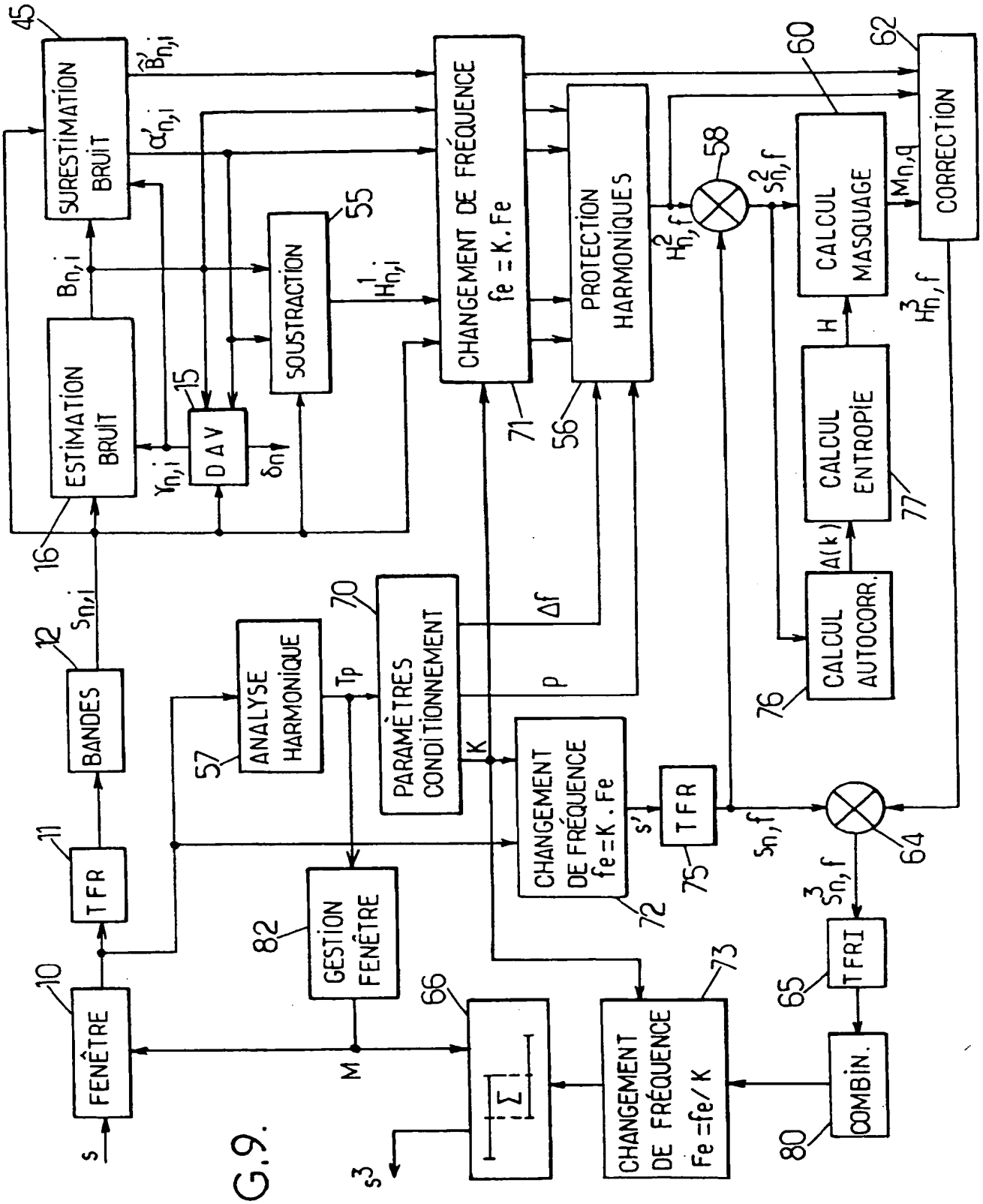

FENÊTRE
TFR
BANDES
ESTIMATION BRUIT
SURESTIMATION BRUIT
DAV
SOUSTRACTION
CHANGEMENT DE FRÉQUENCE fe = K.Fe
PROTECTION HARMONIQUES
ANALYSE HARMONIQUE
PARAMÈTRES CONDITIONNEMENT
GESTION FENÊTRE
CHANGEMENT DE FRÉQUENCE fe=K.Fe
TFR
CALCUL AUTOCORR.
CALCUL ENTROPIE
CALCUL MASQUAGE
CORRECTION
TFRI
COMBIN.
CHANGEMENT DE FRÉQUENCE Fe=fe/K
Σ
s
s3
M
10
11
12
15
16
45
55
56
57
58
60
62
64
65
66
70
71
72
73
75
76
77
80
82

FIG.9.

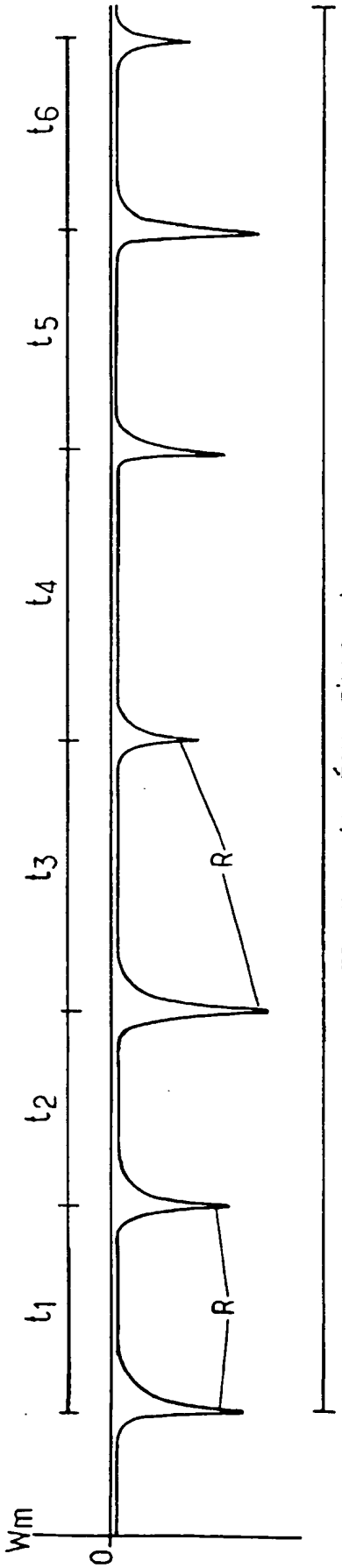

Wm
0
t1
t2
t3
t4
t5
t6
R
R
TRAME n (N ÉCHANTILLONS)
M (≤ N/2) ÉCHANTILLONS SAUVEGARDÉS

FIG.10.

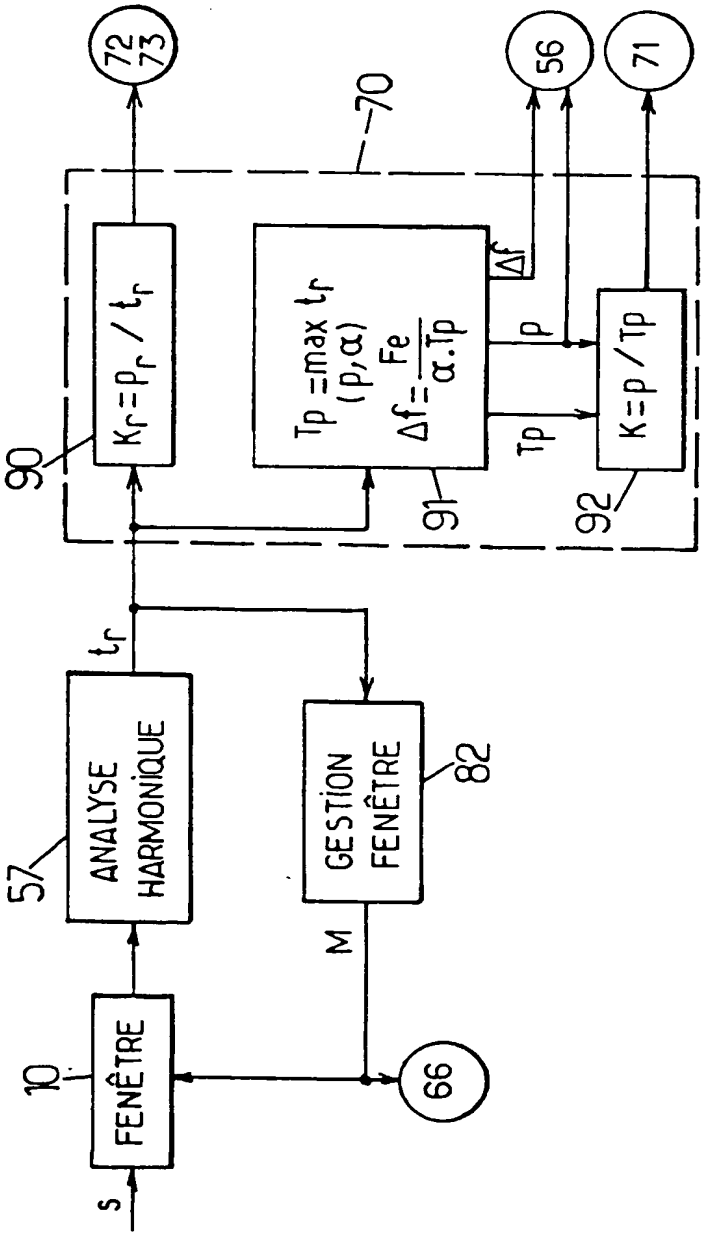

s
FENÊTRE
10
ANALYSE HARMONIQUE
57
GESTION FENÊTRE
82
M
66
tr
70
Kr=pr / tr
90
Tp =max tr (p,α)
Δf=Fe/α.Tp
91
K=p / Tp
92
Δf
p
Tp
72 73
56
71

FIG.11.

REPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
de la
PROPRIETE INDUSTRIELLE

RAPPORT DE RECHERCHE PRELIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

2768547
N° d'enregistrement national

FA 549904
FR 9711643

DOCUMENTS CONSIDERES COMME PERTINENTS		Revendications concernées de la demande examinée
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	
D,A	NANDKUMAR ET AL.: "Speech enhancement based on a new set of auditory constrained parameters" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, SIGNAL PROCESSING (ICASSP 1994), vol. 1, 19 - 22 avril 1994, ADELAIDE, AU, pages 1-4, XP000529345 * alinéa 2 - alinéa 3 * ---	1
D,A	LOCKWOOD ET AL.: "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars" SPEECH COMMUNICATION, vol. 11, no. 2/3, 1 juin 1992, AMSTERDAM, NL, pages 215-228, XP000279184 * alinéa 2.2 * ---	1
A	EP 0 661 821 A (SHARP) 5 juillet 1995 * colonne 6 * ---	1
A	WO 95 02930 A (DOLBY LAB LICENSING) 26 janvier 1995 * page 4, ligne 1 - ligne 18 * -----	1

DOMAINES TECHNIQUES RECHERCHES (Int.Cl.6)

G10L

1

Date d'achèvement de la recherche	Examinateur
18 juin 1998	Lange, J

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : pertinent à l'encontre d'au moins une revendication ou arrière-plan technologique général
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P04C13)